Policy Group: DMP Breakout

Data Plan breakout (Policy Group)

- One size does not fit all (or even most)
- Data management plan needs to be appropriate to the size, scope, duration, and needs of the project or program; aligned to data policy
  - There is always a greater level of detail possible
  - What level of data is appropriate?
  - Ties back to data as an "asset" – investment level
  - Research, resource, reference (litigation)
- Examples of good data management plans at different scales would be useful

Barriers (for data plans)

- Training (need more information workers)
- Different levels of experience progress across agencies in data sharing
- Reuse of data (and the data management plan) generally benefits people other than the data generators
- Scientific disciplines and applications cross multiple agencies
- Different meanings for the same words
  - Loaded connotations in some environments
  - Different divisions (checksums – preservation or quality?)
  - Colliding communities of practice

Barriers (for data plans)

- Overwhelming nature of possible plans
- Need exemplars of data management plans
  - Examples for different types of projects
  - What plans exist for data.gov datasets (dataplan.gov?)
- A variety of independent efforts
- Is “one size doesn’t fit all” used as an excuse for inaction?
- Decisions (policy) not always based on data
- Data are not interoperable

Policy user needs

- Find the data relevant to policy mission
- Understand the uncertainties (quality) of the data
- Timely, complete data sets (I have to make a decision now)
- Understand the data quickly
- Merge disparate data (different domains, different sources of data).
- Unrestricted data as possible (for regulatory transparency and scientific integrity)
- Know that the data exists, where it is, how to get to it

Researchers: DMP Breakout

Full Range Metadata Content Formate
(14 votes)

- Standards
- Supplemental metadata (e.g., crowdsourcing)
- Version
- Discoverability
- Context and dataset
  - Related datasets
  - Producing organizations
  - Data interchange format

Access (Control/Discoverability)
(12 votes)

- Marry with metadata to allow discoverability
- Embargo and sunset periods
- Technical mechanisms for accessing the data
- Pricing information
- Procedures for authorization/access
- Usage restrictions (e.g., noncommercial; copyright; Agency-use-only)
- Security
  - Restrictions and rights
  - Intellectual property rights

Data Quality Management

(8 votes)

- Identify and apply existing measures (objectives, calibration, quality assurance, fitness for use).
  - Scientific data quality
    - Peer review
    - Uncertainties associated with the data
    - Validation processes used
    - Calibration
    - Etc.
  - Data management data quality must each be covered.
    - Data integrity
    - Bit rot consideration

Description
(7 votes)

- High level, editable description as an abstract that can be used as a summary and be helpful for discoverability.

Project Lifecycle Stewardship
(4 votes)

- Considers the future of the data.
- Researchers can be credited for their work (both reports and datasets) in perpetuity.
- Plans in place to store or hand off data (e.g., Agency science data center)
- Personal resource for inquiries of data set

List of Issues that are Impediments

- The elements need a conceptual framework that Agencies can interpret.
- Break the culture of “the data is mine”. Need an improvement to our existing culture.
- Need something like NSF plan so that creators of data get credit.
- Flexibility in release date for appropriate reasons.
- Publisher must answer questions about the data.
- Unfunded mandate.
- Could use a good definition of usability.
- Levels of inheritance & generalization

Science Managers: DMP Breakout

Guiding Principles:

- Data management plan (DMP) complexity should be commensurate with the complexity of the data to be managed.
- DMP template should be organized similar to an agreed-upon data management function model. (Red Team used IWGDD model)
- Identify the appropriate long term repository for data that will be retained.
- Each agency will identify the appropriate approval process for the DMP for that agency

Cultural Challenges:

- Appraisal of data for long term retention is performed in coordination with data producer, not by the data producer

Data Management Function Model for Organizing the Elements

In terms of Data Management Functions:

- Acknowledge that these points may be covered in the research/ project management plan
- Not element is mandatory for every project this list presents all the elements that may be needed

The Elements

Document– PI is responsible and identifies the roles and responsibilities

- **Description:** high level description of the data to be managed
- **Impact**
- **Content:** include metadata, required tools; format; identify what will be stored/saved raw, intermediate, or derived and their respective retention periods, best practice: if standard data structure is applicable it should be used. The description of the size of the data could be important.
- **Workflow**—what is the instrument or processes used to create this data, point to the research management plan if applicable
- **Version control** and change control
- **metadata management,** methods of collection may also be important
- **Provenance**
- **Reference & master data**– identifies the master source and how you generate data from the master data

The Elements

Organize- Data systems stakeholders and potential partners (e.g., other agency, university)

- **Architecture management**—you may need a specific software application, and/or an entirely segregated system, or it may be across multiple systems depending on the access or network requirements of the data. Infrastructure to support data persistent ID management.
- **Identify potential partners**, identify interim roles and responsibilities
- **Data warehouse**

The Elements

Access: PI addresses the audience/relevance of this as appropriate. Other data stakeholders may also need to address this.

- **Usability**, data is usable by others, this is use constraints in the metadata, who can access this data and under what conditions can this data be used
- **Access,** methodologies need to be considered (eg web enabled access), both input of data/information as well as download of data/information
- **Value-added**, data and metadata services should be described, this could include modeling, data mining, data linking,
- **Data sharing**, mechanisms of data sharing including embargo and sunset of data should be included
- **Availability of data** (timeframe of when data is available, is this time critical data?)
- **Strategy of discovery of data**–how will users know where to find your data

The Elements

Protect: PIs need to articulate how to safeguard and/or protect the data. Data stakeholders have an interest in this aspect of data management.

- Operations management– Technical requirements and associated costs need to be included in the planning phase even just an estimated
- **Security,** data security and sensitivities are implied here.
- **Appraisal & disposition**
- **Stewardship/Transfer responsibility**–Identify where/how data will go after primary use, who is the long term custodian, how will it be archived. This is in partnership with PI/data provider and the data stakeholder repository.
- **Quality**– define your QA process to ensure quality data we had a concern about what this is, is it QA of your data or QA of the data as it is being transferred from one location to another

????These are overarching elements?????

- **governance,** does this pertain to the data policy, is this an important element for a DMP it seems like it’s an overriding principle? Perhaps when there are collaborations the governance of the data is important and needs to be articulated?
- **Document and content management**

Issues that need to be considered…

- What are the mechanisms for paying for this?
- MOU across agencies that may have to be adhered to.
- Integration of DMP needs to occur with IT for planning
- Need multiple levels of DMPs agency level strategy, data center, and individual projects—each plan may have different contents. **The DMP needs to identify it's audience, funding agency, data repositories, and this drives the complexity of this document.**
- DMP needs to work in conjunction with the data policy.

Requires better definition…

- **Quality**– define your QA process to ensure quality data we had a concern about what this is, is it QA of your data or QA of the data as it is being transferred from one location to another
- In some ways this is a PI driven document, so **data integration** and the derived benefits of data integration are not discussed, but are important.
  - This is particularly important for large collaborative projects such as those in high energy physics. International collaborations are another issue, in particular where will this international data reside and how will we resolve embargo of data.

Operational Users: DMP Breakout

Key Areas for Operational Users

- Description
- Impact
  - Describe users and uses
  - Will change over the life of the data and needs to be updated
- Content and Format (including metadata)
  - Start the data management early in the life cycle
  - Consider links with other systems to support automatic population and re-use of metadata
  - Connect to Project Mgmt, Facilities Management, Employee Management, systems, etc. (improves interoperability with other systems)
  - Metadata should identify any physical artifacts (core samples, botanical collections, etc.); should some applications be managed as part of the data?

Key Area, Continued

  - Compression and encryption, etc. need to be documented
  - Data Quality Management (link to Data Quality Plan)
  - Consider that the scientist may want the original format back
  - This may differ from a common format for distribution or data mining
- Version Control
  - Documenting process of change control includes what has happened to the data throughout its life

Key Areas, continued

- Access
  - What is the context?
  - Ability to establish links with other objects
  - Conducive to interoperability
  - Discuss access provided by agencies and perhaps commercial vendors and consider future implications
- Appraisal and Disposition
  - Work toward better consultation among agency records management, NARA and agency program people
  - Get input from operational users and producers
- Provenance (term could also be lineage or chain of custody)
- Data Sharing (citations, licensing, user obligations)

Other Discussion Points

- SDM would be different for legacy data management and time forward.
- There may be some decision paths that are embedded in the plan elements
- A process diagram would be helpful and it looks like we could do a single high level diagram
- Can’t have big expectations of what submitters will be willing to do
- The data management plan will be a living/evolving document at different points of the life cycle
  - Responsibility of different roles
  - Needs to go along with the data and the metadata
  - Levels of planning might impact who and what is done out of the plan elements

Other Discussion Points

- Unintended consequences of mashing up different datasets
- The DAMA approach seems to assume structured data while we have been trying to deal with data with a broad definition
- What do you plan to do to increase the value of the data (how you present, how you disseminate, etc.)?

END